From Element-Wise Filtering to Structured Linear Mixing

Brief technical note for time-pressed researchers

Consider first an element-wise multiplication:

$$y_i = c_i x_i.$$

This is the simplest possible linear filter. Each output is controlled by exactly one input value.

Now consider a weighted sum:

$$y_j = \sum_i W_{ji} x_i.$$

This can be viewed as a **sum of several element-wise filtering operations**:

$$y_j = W_{j1}x_1 + W_{j2}x_2 + W_{j3}x_3 + \cdots .$$

Thus the progression is conceptually:

$$\boxed{1 \text{ filter} \rightarrow 2 \text{ filters summed} \rightarrow 3 \text{ filters summed} \rightarrow \cdots \rightarrow \text{many filters summed}}$$

or, in terms of a matrix row,

$$\boxed{y_j = c_1x_1 \rightarrow c_1x_1 + c_2x_2 \rightarrow c_1x_1 + c_2x_2 + c_3x_3 \rightarrow \cdots}$$

Why summing filters weakens the filtering effect

A single element-wise filter has a strong filtering action: one input channel directly determines one output channel.

When several filtered inputs are summed, the output no longer corresponds uniquely to any one input. Information from several channels is combined.

Consequently, the filtering becomes progressively less coordinate-specific and more **mixing-oriented**.

This gives an intuitive continuum:

$$\boxed{\text{strong coordinate filtering} \quad \longrightarrow \quad \text{local information mixing} \quad \longrightarrow \quad \text{global information mixing}}$$

The important point is that a weighted sum should not be thought of as something fundamentally unrelated to filtering. It is **a superposition of elementary filters**.

Connection to observability

This distinction matters for neural networks.

With element-wise filtering,

$$y_i = c_i x_i,$$

information loss is highly explicit. If $c_i = 0$, the corresponding input component is eliminated.

With several filters summed,

$$y_j = \sum_{i \in B_j} c_{ji} x_i,$$

information from multiple coordinates is combined. This can make individual coordinates less directly observable, but can also provide redundancy and allow useful information to survive through alternative combinations.

Thus increasing the number of summed filters changes the character of information flow:

$$\text{selection} \rightarrow \text{local association} \rightarrow \text{distributed representation}.$$

Structured matrices

This provides a useful interpretation of structured neural-network weight matrices.

Instead of going directly from diagonal mappings to fully dense matrices, one can construct matrices from **small groups of summed filters**, optionally interleaved with fast Walsh-Hadamard transforms.

In particular, using multiple width-16 weighted sums provides a potentially useful middle ground:

- much richer information flow than element-wise multiplication;
- substantially more local structure than a fully dense matrix;
- useful local redundancy for associative recall;
- potentially favorable computational structure for SIMD and tensor-core hardware.

The empirical observation that width 16 works particularly well can therefore be interpreted as a possible **filter-to-mixing sweet spot**: each output combines enough elementary filters to obtain useful information integration, without immediately becoming an unrestricted dense mixture.

Key intuition

> **Element-wise multiplication is one filter per output. A weighted sum is several filters whose outputs have been added together. As more filters are summed, the operation moves progressively from explicit filtering toward information mixing.**

This provides a simple way to reason about the trade-off between **filtering, information flow, and observability** in structured neural-network mappings.